Teachers can do it: Scalable identity-based motivation intervention in the classroom☆

Eric Horowitz[a], Nicholas Sorensen[b], Nicholas Yoder[b], Daphna Oyserman[a,*]

[a] University of Southern California, Mind and Society Center, 645 Exposition Blvd, Suite 205, Los Angeles, CA 90089-3333, USA
[b] American Institutes for Research, 10 S. Riverside Plaza, Suite 600, Chicago, IL 60606 USA

A B S T R A C T

Classroom activities aimed at changing students' identity-based motivation (IBM) improve student outcomes by helping students experience school as the path to their adult future identities and their difficulties along the way as signals of the importance of schoolwork. One way to scale these effects would be to have teachers deliver IBM activities. Hence, we asked if, after a brief two-day training, teacher-delivered IBM intervention could meet fidelity standards and if attaining more fidelity matters. We trained all eighth grade teachers in two middle schools (N = 211 students). We used Dane and Schneider's (1988) five-component fidelity model and Durlak and DuPre's (2008) empirically derived threshold and practical maximum standards for fidelity. We found that most classrooms (88%) and students (89%) received IBM intervention at-or-above threshold standard, implying that teacher-based IBM delivery is viable. Moreover, investing in improving fidelity is worthwhile; above-threshold fidelity improved core grade-point-average and reduced risk of course failure.

A R T I C L E   I N F O

Keywords:
Identity-based motivation
Intervention
Motivation
Possible selves
Academic performance
Grade point average

"In the beginning of the year we did a program called *Pathways-to-Success*. It was a program about how we thought of our futures, and if something got in the way, how would we make plans to overcome them. Something [my teacher] always told the class of 2023 was that if it's difficult, it's important. I feel like this is true. In life if you find something difficult like school for instance it is important." (8th grader Middle School Graduation Speech)"

"Thank you for helping us with what will happen later in life...For giving us a pathway to success and now it is our choice to take that path. You helped us find forks that we may have, the decisions we have to make." (8th grader receiving Special Education services, Thank-you Letter to teacher delivering the Pathways-to-Success program)

"This was by far and away the best advisory program we have had and I've been here 10 years. We have had some attempts at it with very little support that have fallen flat on their face. This may be our third or fourth advisory program." (8th grade Science teacher who delivered the *Pathways-to-Success* program)

1. Introduction

Students want to do well in school and go on to college, yet they often fail to attain their high aspirations (Oyserman & Destin, 2010; Oyserman & Lewis, 2017). One way teachers can harness students' high aspirations is to use identity-based motivation to help their students imagine school as the path to their future, generate strategies to succeed on that path, and see obstacles and failures along the way as signaling importance and value (Oyserman et al., 2017; Oyserman, Johnson, & James, 2011). As our opening quotes suggest, both students and teachers appreciate the usefulness of the identity-based motivation (IBM) perspective. Students found the main points of the IBM intervention useful enough to include in graduation speeches and even felt an impulse to write thank you notes to teachers. Indeed, student academic outcomes improve when classroom interventions target identity-based motivation. Analyses of two identity-based motivation interventions revealed significantly improved student academic outcomes at end of school year follow-up (Oyserman, Terry, & Bybee, 2002) and at end of a two-school-year follow-up (Oyserman, Bybee, & Terry, 2006). In these tests of identity-based motivation theory, pairs of college students (Oyserman et al., 2002) or staff holding undergraduate degrees (Oyserman et al., 2006) delivered the intervention. These prior tests were important because they provided support for the robustness of IBM theory by showing significant effects in real-world settings on important academic outcomes (core course grades and risk failing a

---

class) via change in IBM variables. However, they did not test if, after a brief training, teachers can deliver an IBM intervention with sufficient fidelity to have its promised effects. This test is needed if scaling via teacher implementation is to be possible. We take two steps to address this issue in the current paper.

At step one we test the prediction that a brief 2-day in-service training yields sufficient fidelity to likely have effects. At step two we test the prediction that achieving higher fidelity matters for core grade-point average and course failure rates. We focus on a brief 2-day training because teachers are unlikely to be given time for longer training. We focus on sufficient fidelity because a large review suggests that interventions delivered with less than 60% fidelity are unlikely to have their intended effects (Durlak & DuPre, 2008). We focus on implications of higher fidelity because the Durlak and DuPre (2008) review also suggests that practitioners are unlikely to deliver with more than 80% fidelity. Taken together, this range implies that analyses should focus on whether the fidelity threshold of 60% is attained and whether fidelity above 60% and closer to the 80% practical maximum improves targeted outcomes. To situate our results and their implications, we divide the introduction into three sections. First, we review identity-based motivation theory, the evidence that it predicts academic outcomes, its translation to intervention, and the need for testing teacher-led IBM intervention. Second, we describe what fidelity is and how to operationalize it. Third, we specify our research questions.

1.1. Identity-based motivation theory

1.1.1. Operationalization

Identity-based motivation theory is a social psychological theory of motivation and goal pursuit that explains when and in which situations people's identities motivate them to take action towards their own goals (Oyserman et al., 2017; Oyserman, 2015a). Identity-based motivation theory starts with the assumption that people are sensitively attuned to their immediate context and that this shapes identities (dynamic construction). People prefer to act (action-readiness) and make sense of situations (procedural-readiness) in identity-congruent ways—ways consistent with what 'I' and people 'like me' do. However, even though identity (who one was, is, and might become) feels stable, identities are dynamically constructed in context. Dynamic construction means that contexts shape which identities come to mind, what these identities seem to imply for behavior, and how people interpret experienced difficulty. The thing of interest is not that people can change how they regard themselves after sustained effort, but rather the surprisingly large effects that small shifts in context can have on changing how people regard themselves. As detailed next, each component of identity-based motivation (dynamic construction, action-readiness, and procedural-readiness) has been operationalized and its effect on academic performance empirically tested.

1.1.2. Experimental evidence of effects on academic outcomes

In this section, we briefly review experiments documenting effects of identity-based motivation on academic outcomes. First, we consider studies showing that dynamic construction of identity cues action-readiness; readiness to act in ways that fit constructed identity. In one study, researchers subtly shifted what context implied about being a boy (Elmore & Oyserman, 2012). In this study, researchers randomly assigned middle school boys into groups; each group was shown a different graph of accurate statewide census data. One group—the 'men succeed' group—saw a graph showing that men earned more money than women. This graph implied that academic success fits with being a boy. Boys who saw the 'men succeed' graph made more attempts to solve a math task and imagined more school-focused possible identities than boys who saw other graphs. Boys in these other conditions saw graphs that did not mention gender or graphs showing that women are more likely to have graduated high school than men, implying 'women succeed'.

In a second set of studies, also examining the consequences of dynamic construction of identity on action-readiness, researchers subtly shifted what the future self seemed to imply for action by changing the fit between identity and context (Oyserman, Destin, & Novin, 2015). In these studies, researchers randomly assigned students to think about school as a success-likely context in which most students succeed or to think about school a failure-likely context in which most students do not do as well as they hoped. The researchers then asked students to write about their possible identities, with half of the students guided to consider desired possible identities and half of students guided to consider undesired ones. Thus, half of students were led to consider their future self and their current context as fitting together, either because in that context people often fail and their to-be-avoided future self was on their mind, or because in that context people often succeed and their to-be-attained future self was on their mind. The results showed that the action-readiness component of the future self is context-sensitive. That is, students planned to start studying sooner if the way they thought about their possible identities and the way they thought about school fit together. They were more likely to take action after thinking about undesired possible identities while thinking of school as a failure-likely context or after thinking about desired possible identities while thinking of school as a success-likely context.

In other studies examining the link between the dynamic construction of identity and action readiness, researchers used small contextual cues to make studentsexperienced the future self as relevant to the present moment (Destin & Oyserman, 2009; Destin, 2017; Landau, Oyserman, Keefer, & Smith, 2014; Nurra & Oyserman, 2018). Nurra and Oyserman (2018) randomized students to consider their adult future self as occurring soon or occurring later, as connected to their current self or as distinct from their current self. Experiencing one's adult self as near and connected is consequential for behavior. Across studies, if researchers led students to experience their adult and present selves as connected, students worked harder on current assignments, focused more on boring tasks, and actually attained better core course grades by the end of the semester.

Destin and colleagues (Destin & Oyserman, 2009; Destin, 2017) randomized middle school students to either learn about need-based financial aid (open path) or to estimate the cost of college and report how they planned to cover this cost (closed path). Students who learned that income is not a barrier had significantly higher school engagement compared to students who were asked to consider the cost of college and how they would pay for it. Students who were asked to consider college cost and how they would pay for college seemed to infer that cost was a barrier and hence college was not likely for them, making hard work in eighth grade feel like a pointless endeavor. Landau and colleagues (2014) randomized students to either think about their academic possible identities in the context of an image that implied action (a path) or one that did not (a container). Students led to list their academic possible identities on an image of a path rather than an image of a container were more engaged with their schoolwork. Across studies, these students were more likely to seek out academic help, worked harder on current assignments, planned to study more for an upcoming quiz, and actually performed better on the quiz.

In addition to cuing readiness to act, dynamic construction of identity also cues procedural-readiness—that is, how one makes sense of experienced ease and difficulty with schoolwork as implying something about oneself. Interpretation of experienced difficulty matters for downstream behavior and for identity. Across studies, once students considered that experienced difficulty might be a sign that schoolwork is important, they saw academics as more central to their identity (Aelenei, Lewis, Oyserman, 2017; Oyserman, Elmore, Novin, Fisher, & Smith, 2018; Smith & Oyserman, 2015) and did better on a variety of school tasks (Elmore, Oyserman, Smith, & Novin, 2016; Oyserman et al., 2018; Smith & Oyserman, 2015). Students are also more likely to endorse the idea that difficulty means importance if they experience fit between identity and context, as we previously described. That is, when

students think about their undesired future self and think of school as a context in which failures are likely, then they are more likely to endorse a difficulty-as-importance mindset. The reverse is also true—students are more likely to endorse a difficulty-as-importance mindset when they think about their desired future self and think of school as a context in which success is likely (Oyserman et al., 2015).

1.1.3. Translating experiments to intervention

Evidence to date shows that identity-based motivation intervention can matter by changing student engagement, effort, and grades. The *School-to-Jobs* intervention translated the three core components of identity-based motivation (dynamic construction, action-readiness, and procedural-readiness) into a set of activities (Oyserman et al., 2002; Oyserman et al., 2006). Randomized control trial evaluation showed that the intervention changed elements of identity-based motivation by the end of eighth grade (Oyserman et al., 2002) and these changes mediated changes in core course grades and course failures by the end of 9th grade (Oyserman et al., 2006). Core course grades and course failures are important metrics because core course grades and course failure by 9th-grade increase the likelihood of on-time high school graduation (Allensworth & Easton, 2005), and failing even a single course reduces the likelihood of high school graduation (Allensworth & Easton, 2007).

In developing and delivering the intervention, the intention was that each activity provides students with a different concrete experience of one or more of the components of identity-based motivation in a way that made it likely that students would internalize the core ideas. To do so, activities allowed students to discover and experience each component of identity-based motivation on their own, rather than to be told about identity-based motivation by their instructor, thereby reducing the chances of reactance (Brehm & Brehm, 2013; Elmore et al., 2016) and increasing the likelihood of deep processing of messages (Petty & Cacioppo, 1986). Activities were group-based rather than occurring alone to increase the chances that student social identities—as students, boys or girls, or members of their racial-ethnic group—were cued as 'we do school', increasing the chances that an identity-based motivation cycle would ensue (Oyserman, 2007). When delivered and received as intended, participating students should experience change in their identity-based motivation, and this should improve academic outcomes. Specifically, students should have more school-focused possible identities and the strategies to work on these identities, be more likely to see difficulty with schoolwork as implying its importance, and be less likely to see difficulty with schoolwork as implying that schoolwork is 'not for them.' They should be more likely to see school as the path to their adult future self, see school-focused possible identities as congruent with important social identities, and they should be more likely to experience their adult future self as relevant to their current schoolwork. Over time, these changes should result in better school grades and less likelihood of failing classes. That is what researchers found (Oyserman et al., 2006).

1.1.4. Moving from trainers to teachers

Initial tests of the identity-based motivation intervention showed that researchers could train undergraduates or people with undergraduate degrees to go into schools and deliver the intervention as intended after 40 hours of training. However, these tests did not use teachers. Using people who came and left rather than teachers provided a clean test of the theory, but for practical application, teachers are needed so that IBM intervention can be rooted in a school system. Outside trainers may come and go, but teachers can maintain an intervention over time. Teachers' time is limited and hence a test in which a brief training is provided to teachers and fidelity is assessed is a first step in addressing whether IBM interventions might scale through teacher-delivery. We chose a 2-day test as the briefest likely sufficient training for teachers to be able deliver and their students receive an identity-based motivation intervention as intended. This combination

of teacher delivery-as-intended and student receipt-as-intended is termed *implementation fidelity*.

Ascertaining implementation fidelity is critical for both theory-testing and pragmatic reasons, as detailed next. From a theory-testing perspective, without knowledge of implementation fidelity it is not possible to know whether any changes after intervention are due to the theory on which the intervention is based. At the same time, pragmatically, as we outline next, without implementation fidelity, interventions are empirically less likely to have their intended longer-term effects (Century, Rudnick & Freeman, 2010; Durlak & DuPre, 2008).

Program differentiation is the aspect of implementation fidelity related to theory testing. It is an assessment of whether the ingredients of an intervention are operationalizations of the theoretical process model or theory of change the intervention is based on. If the program uses ingredients that are not part of the theoretical process model or if the same ingredients (perhaps differently labeled) are also in other programs (or the control group), program differentiation is low. There is not much point in delivering multiple interventions with different names that deliver the same intervention ingredients or in delivering an intervention with ingredients not linked to an empirically validated process model of change. Program differentiation assessment might be obtained from a school district, principal, teacher, or researcher, and is at the level of the intervention itself—the program either is differentiated or is not differentiated from other programs. In the case of identity-based motivation intervention, the question would be whether other programs in the school use IBM theory or IBM ingredients otherwise labeled. Our principals and teachers concluded that they were not already delivering an identity-based motivation intervention or even another socio-emotional learning (SEL) program.

1.2. Implementation fidelity entails fidelity of delivery and of receipt

Aside from program differentiation, the other components of implementation fidelity are *dosage*, *adherence*, *quality of delivery*, *student responsiveness* and *fidelity-of-receipt* (Dane & Schneider 1998; see also Bellg et al., 2004; Dusenbury, Brannigan, Falco, & Hansen, 2003; Crosse et al., 2011; Mowbray, Holter, Teague, & Bybee, 2003; O'Donnell, 2008; Resnick et al., 2005). To be useful, implementation fidelity operationalization (O'Donnell, 2008) and report (Hulleman & Cordray, 2009) should fit the intervention itself. For example, in the case of classroom-level delivery, it is reasonable to expect classroom-level and student-level variation—some classrooms and some students will experience more implementation fidelity (experience faithful delivery and message uptake) than others. Rather than being independent, the components of fidelity are best understood as interdependent building blocks that scaffold and support each other.

At the base of fidelity are *dosage*, the timing and number of sessions delivered compared to plan, and *adherence*, the extent that each activity in each session is delivered in the sequence and as the manual describes it. *Dosage* and *adherence* scaffold *quality of delivery* and *student responsiveness*. *Quality of delivery* entails teacher-managed session 'feel,' which teachers produce via classroom and student-level emotional and organizational support and behavior management and via their structuring of delivery of take-home points. High quality delivery entails students experiencing take-home points as emerging from themselves rather than from teachers and as easy to process and hence true, rather than as emerging from teachers, difficult to process, and hence not necessarily true. *Student responsiveness* entails student response to *adherence* and *quality of delivery*. When teachers deliver (*dosage*) the correct content (*adherence*) in the correct way (*quality of delivery*), their students should respond with engaged attention and productivity (*student responsiveness*) and hence internalize the take-home points (*fidelity of receipt*). Given the interrelatedness of each element conceptually, teasing apart these elements of fidelity analytically would require randomizing teachers to deliver varying doses or to adhere in varying levels or to separately deliver take-home points with varying quality.

Researchers can use classroom observation to obtain classroom-level ratings of how much of intended intervention intensity and duration was delivered (*dosage*), how much delivery followed protocol (*adherence*), and how much participants responded as intended (*student responsiveness*). Researchers can use classroom observation and student reports to obtain *quality of delivery* ratings. Researchers can use student report to assess the extent that participants have received and understood take-home points (*fidelity of receipt*).

1.2.1. Reasonable expectations for fidelity

Having defined fidelity, the next questions are how much fidelity can be expected, and how much is sufficient for an intervention to have its desired effects. To address these questions, Durlak and DuPre (2008) reviewed the meta-analytic literature on interventions delivered in real world settings by non-researchers, adding 59 additional studies they found that were not included in the prior meta-analyses. They asked whether this literature pointed to a threshold at which, on average, an intervention yields its desired effects and whether this literature provided guidance into how much fidelity non-researchers could be expected to produce. They concluded that studies rarely show intended effects unless fidelity reaches or surpasses a 60% threshold and that non-researchers rarely implement with greater than 80% fidelity. These findings have a number of implications. First, fidelity researchers should expect that a successful training would yield fidelity between 60% and 80%. Second, researchers should test whether moving from 60% to 80% fidelity increases likelihood of attaining intended impacts, and if so, if the increase is linear or looks more like a step function. If the increase looks more like a step function, researchers should test where the step-up occurs.

1.2.2. The 60% fidelity threshold in educational research

We examined school and education-focused evaluation research published since the Durlak and DuPre (2008) threshold and practical maximum estimates were published. We found that Durlak and DuPre's (2008) 60% fidelity threshold is used across an array of community-based and school-based intervention evaluations to document that sufficient fidelity is attained. For example, Fagan, Hanson, Hawkins, and Arthur (2009) used this threshold in a 12-community evaluation of Communities that Care, a system for community partners to utilize prevention science. Riordan, Lacireno-Paquet, Shakman, Bocala, and Chang (2015) used the 60% threshold as a marker of necessary fidelity in examining the implementation of a teacher evaluation system in 15 schools. Bloomquist and colleagues (2013) used the 60% threshold as an indicator of sufficient fidelity in the implementation of an intervention that aimed to reduce conduct problems in 27 elementary schools. Lindsay, Davis, Stephan, and Proger (2017) used this threshold in evaluating a college readiness program in 25 schools. At the preschool level, the 60% threshold is used in testing implementation of the preschool-based Early Literacy and Learning Model in 28 preschool classrooms (Preschool Curriculum Evaluation Research Consortium, 2008). The 60% threshold is also used by evaluators of Life Skills Training, a widely-used school-based substance use prevention intervention evaluated in over 30 peer-reviewed studies of programs in over 300 schools involving 20,000 students (e.g. Velasco, Griffin, Botvin, Celata, & Lombardia, 2017; Botvin & Griffin, 2015; Botvin, Baker, Dusenbury, Botvin, Diaz, 1995).

While Durlak and DuPre (2008) describe a linear relationship between fidelity and outcomes, they did not separately examine whether outcomes improve as program fidelity shifts from the 60% threshold to the 80% practical maximum. We did not find other papers addressing this question either so it is not clear if investing in fidelity beyond the 60% threshold matters, and if so, if improved outcomes are linearly associated with improved fidelity.

1.3. Research questions

Our review of the fidelity literature led us to two research questions. First, can teachers deliver and students receive an identity-based motivation intervention at or above the 60% fidelity threshold? Second, does moving beyond the 60% fidelity threshold matter for student core grade-point average and for their likelihood of course failure? We used the mean of the five components of implementation fidelity (*dosage*, *adherence*, *student responsiveness*, *quality of delivery*, *fidelity of receipt*) that could vary at the classroom and student levels to test our first research question, and the relationship between implementation fidelity and school grades and course failure rates to test our second research question. To further understand the nature of our fidelity effects, we also compared one aspect of teacher quality—teacher-driven classroom climate—in the teacher's subject classroom (while teaching math, science, history, or language arts) with the teacher's classroom climate score while delivering the identity-based motivation intervention. This comparison allowed us to begin to address whether the quality aspect of fidelity was capturing what the teacher did in the identity-based motivation intervention or something more general about the teacher. Finally, because our goal was to learn how to improve fidelity, we examined teacher responses to our queries about ways to improve usability and feasibility and looked carefully at the classroom experience to build and revise for future implementations.

2. Materials and methods

2.1. Sample

In the first year of our grant, all eighth-grade teachers (eight classrooms) in two Chicago K-8 public schools and their full cohort of eighth graders (N = 211, 50% female, 93% nonwhite, 94% free or reduced lunch eligible) participated in the intervention.[1] Classroom size ranged from 25 to 31 students except for one pullout classroom, with 12 students receiving special education services.[2] Three teachers were female, seven were white; their main subjects were Math (two teachers), Science (two teachers), English (two teachers), Special Education (one teacher), and History (one teacher). All teachers had three or more years of teaching experience, making them similar to the statewide average—88% of teachers in Illinois have three or more years of experience. The schools themselves were at or below state average in terms of their standardized test scores for 8th grade. Statewide, 40% of students scored in the range labeled "met or exceeded expectations" in English; in our schools, the percentages were 37% and 19%. Statewide, 32% of students met or exceeded expectations in Math; in our schools, the percentages were 35% and 26%.

Analyses describing fidelity (mean of the five fidelity components, *n* = 184) employ listwise deletion for students missing student-level data on fidelity. Student-level data were missing if parents refused to consent to data collection (*n* = 4) or if the data were missing even though parents had consented—presumably because the student was absent the day of data collection (*n* = 23). Analyses describing how fidelity affects subsequent course grades (*n* = 209) employ listwise deletion for missing data on 8th grade academic outcomes. Only 1% of 8th grade academic data are missing as a result of students leaving the district, an additional five students were missing 7th grade academic data, 19 were missing student-level fidelity data, and 7 students were missing both 7th grade academic data and student-level fidelity data. To preserve the analytic sample in our regression models, we imputed

---

[1] Students were from a variety of racial-ethnic backgrounds: 68% were of Hispanic background, 16% were of Asian background, and 9% were of African American background. The remaining students were categorized as having White (7%) or multiracial-multi ethnic backgrounds (1%).

[2] We present results that exclude this classroom in our Supplemental Materials. Inclusion or exclusion of these data does not change our results.

missing data and used a dummy variable to adjust for missingness in covariates. Specifically, we imputed 7th grade academic data by assigning students the average score in their classroom and imputed missing student-level fidelity data by assigning students their classroom average.

2.2. Procedure

We maintained continuity in training; Oyserman, who led the *School-to-Jobs* training, also led training and weekly check-in calls (Oyserman et al., 2006). The implementation manual and training highlighted how to deliver with high quality. This included seven components: (1) scaffolding activities to both be personal and generate a sense of group norms of engagement, (2) keeping a good pace, (3) creating a positive emotional climate, (4) being well organized, (5) delivering the content in a way that appropriately evokes participation, (6) delivering the content clearly in a way that facilitates student experience of each session as naturally unfolding and building on prior lessons, and (7) delivering the content clearly in a way that facilitates student experience of take-home points as student-generated, not teacher-taught. When delivered with quality, students should experience ease in the concrete activities; feel that they, their classmates, and teacher are trustworthy, warm and knowledgeable; and that together they generate useful knowledge.

We made a number of decisions based on our goal of enhancing scalability. First, we made the *School-to-Jobs* intervention implementation manual applicable to teachers—instead of referring to two trainers, all instructions referred to a single teacher—and one activity that the trainers took two sessions to deliver was consolidated into a single session. Second, we used a 2-day (including breaks for breakfast and lunch) abbreviated form of the 5-day training Oyserman provided to trainers in the *School-to-Jobs* intervention. Training took place in a classroom in each school on two consecutive days in August prior to the September start of the school year. Third, we allowed schools some variability in pace of delivery—*School-to-Jobs* was delivered twice weekly but we allowed each school to choose if they would deliver twice per week or once per week. Fourth, we allowed schools some variability in time of day and where the weekly check-in during implementation would occur—either in school with all teachers physically present, or outside of school with teachers calling in. Finally, we followed the original model, which suggested that the intervention be named something that felt meaningful in context (e.g. Oyserman, 2015b). In consultation with participating teachers and schools, we named the teacher-led intervention *Pathways-to-Success,* or *Pathways* for short.

Teachers implemented the intervention during a designated advisory period during the school day, insuring that each student was assigned a single teacher. One school chose bi-weekly delivery for six weeks. In this school, teachers gathered together after school for a video call and finished delivery by Halloween (the end of October). Students receiving special education services participated in the intervention in a separate pull out classroom. The other school chose weekly delivery for twelve weeks. In this school, teachers called into the weekly call in the evening from their own homes and finished delivery prior to Thanksgiving (end of November). Students receiving special education services participated in their regular classrooms.

We maintained continuity in measurement of fidelity: we did not change fidelity materials from the original *School-to-Jobs* with one exception. In *School-to-Jobs*, we used an intervention-specific measure of trainer quality of delivery coded by observers in addition to student-level report, as detailed in Oyserman (2015b). As it turns out, this measure was similar to a widely used standardized measure of teacher instructional quality , the Classroom Assessment Scoring System-Secondary (CLASS-S; Pianta, Hamre, Hayes, Mintz, & LaParo, 2008; Allen, Gregory, Mikami, Lun, Hamre, & Pianta, 2011). Using the CLASS-S requires a two-day training for initial certification in coding and an

annual recertification test in its use—both of our coders met these requirements. Given our goal of communicating with schools, we replaced our prior observer-based measure of quality of delivery with the CLASS-S since we assumed schools would respond more positively to information based in part on a familiar metric.

To obtain high quality data to assess fidelity we video-recorded each intervention session of each teacher. Pragmatically, this meant that immediately preceding each session an American Institutes of Research (AIR) staff member came into the classroom and positioned and turned on an iPad on a tripod. At the end of each session the staff member came to collect the equipment; video was then loaded onto the AIR secure server for coding. We also obtained student reports at the end of the intervention using an online questionnaire. Chicago Public School District provided school grades for all students as part of a master data sharing agreement with AIR. We computed fidelity as the mean of *dosage*, *adherence*, *student responsiveness*, *quality of delivery*, and *fidelity of receipt* to test whether teachers could deliver at or above the 60% threshold and whether moving beyond threshold mattered for student core grade point average and likelihood of course failure.

Immediately following the final session of the *Pathways* intervention, a member of the AIR team interviewed each teacher asking a range of questions pertaining to usability and feasibility of delivering the intervention (e.g. "How did the resources provided by the *Pathways-to-Success* program help you implement the program given your other responsibilities and time commitments?"; the full set of questions is located in Supplemental Materials, Section 1). The goal of this interview was to help us learn what obstacles teachers encountered during implementation regarding preparing for and delivering each session of the intervention. We then used their feedback to guide our plans for improvements aimed toward increasing scalability, as described in our discussion of practical implications in section 4.3.

The week after the final session of the *Pathways* intervention, students completed a brief end-of-intervention survey focused on fidelity of receipt and their perceptions of aspects of quality of delivery. This survey included a brief set of parallel questions about an element of teaching quality (*teacher-driven classroom climate*) for the student's math, science, English, and history teachers to allow for analysis of whether the training itself or aspects of teachers generally influenced *quality of delivery* in *Pathways* (the full student end-of-intervention questionnaire is located in Supplemental Materials, Tables S2.13, S2.14, and S2.16).

2.3. The intervention

The full intervention manual that the teachers in this study used is published (Oyserman, 2015b). As an overview, Table 1 provides each session's thumbnail sketch, take-home point, and core identity-based motivation active ingredient.

2.4. Consent

The school district approved our human subjects' protocol. We included in our fidelity analyses only students with parental consent for survey collection; almost all (98%) provided it. To reduce teacher burden and ensure that paperwork was complete, an AIR staff member handed out and collected parental consent forms and each student was given two movie tickets after returning a form, regardless of whether parents provided or withheld consent. All teachers signed consent forms for video recording. Prior to coding, faces of students without parental consent for video recording were blurred out.

2.5. Fidelity

We assessed *dosage* (see Supplemental Materials, Tables S2.1-S2.11*)*, *adherence* (see Supplemental Materials, Tables S2.1-S2.11), and *student responsiveness* (see Supplemental Materials, Tables S2.1-S2.11;

Table 1
Thumbnail sketch of each Pathways session, take-home point and activated IBM ingredients.

Session	Classroom activity flow	Take home point	Activated IBM constructs
1. Setting the Stage & Introductions	Students are paired up and briefly interview one another on the skills or ability they each have that will help them complete the school year successfully (e.g., "well organized," "positive attitude"). Then each student introduces his or her interview partner in terms of these skills.	We all care about school, and we have a skill or ability to work on our 'successful in school' possible self.	DC, AR
2. Adult Images	Students pick photographs that fit their adult "images." Images of what their adulthood will be like. Photographs include the four domains of adulthood: material lifestyle (e.g., homes), job (e.g., working at various jobs), relationships (e.g., family, friends), and community engagement (e.g., volunteering, voting). Photographs include both genders and match the racial-ethnic makeup of the school. Domains of adulthood emerge from clustering student responses and having students name these clusters.	We all have images of ourselves as adults in the far future.	DC
3. Positive and Negative Forces	Students draw or write about positive and negative forces—people or things that energize them to work toward their possible identities by showing what to do or what not to do.	Everyone faces obstacles and difficulties; positive and negative forces help by laying out paths to take or avoid and ways to handle obstacles and examples of what not to do.	DC, PR
4. Timelines	Students draw timelines into the future, including forks in the road and obstacles. Since students start with the present, all timelines involve school.	Present and future are linked on a path. There are choices and obstacles. Current actions set up which futures are possible. Obstacles must be gotten around to get back on path.	DC, PR
5. Action Goals	Students write action goals, linking next year and adult possible selves with actions they can take right away in a specific time and place to concretize the plan. They do this using an easy to recall formula (because… I will… when…).	We have some control over possible selves, but not our hopes and dreams -control happens when we link the future with the present through specific 'action paths' ways to move to the far future by working now to attain near future goals.	DC, AR
6. Possible selves and strategies	Students map out their expected and to-be-avoided possible selves and strategies for next year on a pathways board.	Strategies are actions you are taking now or could take to become your next year possible self.	DC, AR
7. Pathways to the Future	Students complete pathways boards to concretize the link between current strategies for action, next year possible selves, and adult possible selves.	Strategies I'm doing (or could be doing) now to get to my next year possible self also help me get to my adult possible self.	DC, AR
8. Puzzles	Students break down problems that seem impossible and use strategies to solve them.	Difficult things can seem impossible, not worth your time; but difficulty can be a signal of importance. When something feels really difficult, you can use a strategy like breaking it down into parts.	PR
9. Solving Everyday Problems	Students write about a school problem they have. They use the IBM skills they have learned to consider how to break the problem down.	Everyday problems can be broken down using the *Pathways* skills you already have. You can consider how it relates to your adult possible self and to your next year possible self, you can consider what your positive and negative forces are in this situation, you can consider what is the choice point or obstacle in this situation and you can ask what are your strategies to get around it.	PR
10. Everyday problems: High school and beyond	Students brainstorm what is needed to finish high school, and see the requirements. Then they brainstorm what is needed to get to college, and see the requirements for that.	You can identify the steps to get from 8th grade to graduating high school.	DC, AR, PR
11. Wrapping up and Moving Forward	Students name the *Pathways* sessions, what each was about, what they liked and what they would improve. This provides a bird's eye view of the full set of activities, closure at ending *Pathways*, and reinforcement of the three IBM components.	What I do now matters for attaining my next year and adult possible selves. Possible selves that are linked to strategies and to a time and a place of action become action goals. There are forks (choices) and roadblocks (failures) along the way. It will be difficult and may feel impossible, but asking questions helps break down what I need to find out and helps me connect to others – positive forces – as well as learn from negative forces what not to do.	DC, AR, PR

Note: DC = Dynamic construction, AR = Action-readiness, PR = Procedural-readiness.

17

Did the pace, repetition and clarity together converge to create a fluent experience (must be true)?		
Message feels untrue.	Sometimes feels true, sometimes feels untrue.	Must be true

Fig. 1. Quality of Delivery Rating-Scale for Take Home Point Fluency.

and Supplemental Materials Table S2.12)[3] from video records of sessions. We assessed *quality of delivery* from session video records and from end-of-intervention student report (see Fig. 1, Supplemental Materials Tables S2.12-S2.14) obtained the week after the intervention ended. We assessed *fidelity of receipt* with end-of-intervention student report (see Supplemental Materials, Table S2.15). As detailed below, our coders coded video of each session for each teacher using a structured protocol and we scaled student-report data.

2.5.1. Reliability of Video-based coding

The third author coded all elements of fidelity obtained from video records using the structured protocols described next. We assessed inter-rater agreement in two ways. First, to assess inter-rater agreement in our structured protocol, we had another AIR staff member who was also CLASS-S certified code the full protocol in nine randomly selected video records. To assess our CLASS-S inter-rater agreement, we had the AIR staff member code CLASS-S in thireen randomly selected video records. We used two metrics for coding inter-rater agreement (reliability): percent agreement and Cohen's Kappa (Fleiss & Cohen, 1973). Cohen's Kappa is useful given that coding is categorical. For ease of comparison, we report average reliability when we coded multiple measures for a given fidelity component. To provide some rule of thumb for Cohen's Kappa, Landis and Koch (1977) suggest that scores in the .4 to .6 range represent moderate agreement, while scores in the .61 to .8 range represent substantial agreement. Taken together, coder agreement is sufficient: *Dosage* (85% agreement, Cohen's Kappa = .59), *adherence* (78% agreement, Cohen's Kappa = .55), *student responsiveness* (88% agreement, Cohen's Kappa = .75), *quality of delivery* (75% agreement, Cohen's Kappa = .60).

2.5.2. Computing dosage fidelity

Dosage (α = .84[4]) is a mean score across the 11 sessions. In each session, *dosage* had two components: *implementation* and *task percentage*. *Implementation* was the percentage of sessions that teachers implemented (counted from the video). Because all teachers implemented each session, each teacher scored 100% on *implementation*. We calculated *Task Percentage* by dividing the number of tasks the teacher actually facilitated by the number that were to be facilitated in each session and multiplying by 100. To obtain this number, the observer watched the video of each session and marked with a check if the task occurred or not. We used a checklist to measure these tasks (Supplemental Materials, Tables S2.1-S2.11; Oyserman, 2015b). For ease, Table 2 shows the checklist for Session 2 of the intervention; the first column is what was counted for *task percentage*. The number of tasks in each session varied from 9 to 20.

2.5.3. Computing adherence fidelity

Adherence (α = .95) is a mean score across the 11 sessions. In each session, *adherence* was the count of the number of specific teacher actions the teacher actually took in the session divided by the number that was to be taken in each session and multiplied by 100. We used a checklist to measure these actions (Supplemental Materials, Tables S2.1-S2.11; Oyserman, 2015b). The observer watched the video record of each session and marked with a check if the action occurred or not. For ease, Table 2 shows the checklist for Session 2 of the intervention; the second column is what was counted for *adherence* percentage. The number of actions in each session varied from 15 to 39.

2.5.4. Computing quality of delivery fidelity

Quality of Delivery (α = .74) is a mean score across components. In each session, we computed from two sources and each had multiple components: observer coding of each session's video record and end-of-intervention student report (Fig. 1; Supplemental Materials, Tables S2.12–S2.14 respectively). We calculated a *quality of delivery* score for each student in two steps. At step one we obtained a percentage of total possible points in each metric. At step two we obtained a mean percentage across the metrics.

For observer-based elements of quality, the observer, certified in the CLASS-S, watched each session video and rated 11 dimensions of quality twice per session. The 11 dimensions were organized in three domains: *Emotional Support, Organizational Support,* and *Instructional Support*. Emotional Support includes three dimensions (Positive Climate, Teacher Sensitivity, and Regard for Adolescent Perspectives). Organizational Support includes three dimensions (Negative Climate [reverse coded], Productivity, and Behavior Management). Instructional Support includes five dimensions (Instructional Learning Formats, Content Understanding, Analysis and Problem Solving, Feedback, and Classroom Discussions). To code, the observer stopped the video at approximately the 20-minute mark and the 40-minute mark (or end) of the session. The two scores on each dimension were averaged to a lesson score for each dimension; lesson scores were averaged to obtain a final score for each teacher. Coding was on a 7-point scale from 1 = *not all characteristic* to 7 = *highly characteristic* (negative items reverse-coded). Dimensions were coded by observing classroom interactions. Each dimension has a unique scoring rubric (Pianta et al., 2008) and specific behavioral indicators associated with low (1–2) mid (3–5), and high (6–7) range scores.

For example, the Productivity dimension's behavioral indicators include maximizing learning time, routines, transitions, and preparation. In the low range, teachers provide few tasks for students to complete, the class is disorganized and the students do not appear what to do, the students spend a significant time in transitions, and the teacher is not prepared for the session. In the mid-range, teachers provide tasks for students to complete the majority of the time, but the learning is sometimes disrupted or there are inefficiencies in managerial tasks. There are times of uncertainty and there are some inefficiencies in transitions but mostly there are classroom routines. The teacher is mostly prepared but has some last minute preparations. In the high-range, the students have tasks to complete, are comfortable with the routines, transitions are efficient, and teachers are prepared to deliver the lesson. Interested readers can find the general CLASS-S manual scoring rubric in Table S1.12 in the Supplemental Materials.

At the end of each session, observers read the session take-home point and coded how *Fully* (1, third column) and how *Fluently* (Fig. 1) each session's take-home point was conveyed. For the *Fully* component, take-home scores ranged from 0 to 2—the take-home point was: 0 = not evoked at all by activities, 1 = partially evoked but with unclear or inconsistent framing, 2 = clearly and consistently evoked with

---

[3] Table S1.14 provides the CLASS-S General Scoring Rubric. CLASS-S is a proprietary instrument; hence we cannot include the full manual. www.teachstone.com provides more information about the CLASS instruments.

[4] Alpha reliability used only on the *task percentage* component because the *implementation* component was invariant across teachers (all implemented all sessions) and hence cannot be used to calculate alpha.

Table 2
Checklist for coding dosage, adherence, and student responsiveness in session 2 (adult images).

Task	Y	N	Detailed teacher behavior	Y	N	Student behavior	Y	N
Agenda hung								
Opening								
Welcome			Greet participants and latecomers			Greet trainer		
			Say today is session 2, adult images					
Last session			Ask for what happened last session and why			Share ideas (Learned names about each other, expectations, concerns, games as a team, adding and building on each others' skills)		
			Reinforce student participation (Why: people have lots of different skills that will help them succeed)					
Bridging			Teacher bridges last session and this session (last session we focused on skills and abilities to succeed in school, today we want to look towards the future and the adults we want to become)			Listen		
Images								
Introduce the concept of adult images			Explain task – choosing pictures that represent images of yourself as an adult. Each to pick 3 to 5 pictures, what do they mean for you and when these will be true of you, afterwards share			Listen		
Create personal images			Make instructions clear/Ask for questions			Ask questions/Clarifies directions		
			Have participants begin			Move around room, picking pictures		
			Mingle – check for understanding					
			Have everyone rejoin circle					
Share			Explain task – show 1 picture and explain to group, while group listens and pays attention			Participate		
			Write participant responses on newsprint, clustering by themes			Listen		
Domains of adulthood								
Highlight various domains			Explain task – participant to call out what they thought was similar about everyone's adult images			Share ideas		
Reinforce personal competence in noticing connections, ability to contribute to the in group			Highlight themes that emerge (e.g., jobs, family, friends, community involvement, life style; trainer need only mention domains that did emerge)			Listen		
Next session and goodbyes			Summary Statement: adult images can be about jobs, family, friends, community involvement, and lifestyle (only those group brought up or implied) (adult images + repeat themes)			Listen		
			Connecting statement: next session we'll identify models and forces that help us work on those adult images that are goals					
Completed necessary components of session in appropriate time								

concepts connected to student-generated examples. We coded *Fluency* as 0 (thumbs down) if the session pace, repetition , and clarity of delivery together converged to create a disfluent experience in which the take-home point did not ring true. We coded *Fluency* as 1 (sideways thumb) if pace, repetition , and clarity converged to create some points in which the take-home rang true and some points in which it rang false. We coded *Fluency* as 2 (thumbs up) if session pace, repetition, and clarity converged to create a sense that the take-home point must be true.

At the end-of-intervention survey, students provided their quality ratings on four scales that assess classroom quality on dimensions compatible with the CLASS-S. Students rated their *teacher's sensitivity* (1 = not at all, 5 = a lot; α = .75): how often the teacher understood their problems, listened to their comments, negatively criticized their ideas (reverse-coded), used specific examples, gave everyone an equal chance to participate, and gave students the chance to answer one another's questions. Students rated two aspects of classroom positive climate (teacher-driven and classmate-driven). Students rated the *teacher-driven classroom climate* (1 = strongly disagree, 5 = strongly agree; α = .84): how enthusiastic, warm, clear, and knowledgeable their teacher was. Students rated their *classmate-driven classroom climate* (1 = strongly disagree, 5 = strongly agree; α = .85): how enthusiastic, warm, clear, and knowledgeable their classmates were. Finally, students rated *classroom regard for adolescent perspectives* (1 = strongly disagree, 5 = strongly agree; α = .67): how often they felt comfortable asking questions, they could trust others to listen to what they had to say, others shared their experiences and difficulties working toward their futures, it seemed that other students had the same problems they did, what they talked about was relevant to them, and they felt concerned they would be negatively criticized by another group member (reverse-coded). The exact wording of each item is Table S2.13 in Supplemental Materials.

2.5.5. Computing student responsiveness fidelity

Student responsiveness (α = .82) is a mean score across the 11 sessions of the two components scored in each session: *Student behavior*, as measured by the checklist, and the *Student Engagement* dimension of the CLASS-S. Observers watched the video of each session and as the session unfolded, they marked with a check if an expected student response occurred or not following the original School-to-Jobs checklist of student responses in each session (Supplemental Materials Table S2.1-S2.11; Oyserman, 2015b). For ease, Table 2 shows the checklist for Session 2 of the intervention, the third column is student responsiveness. The number of responses expected varied by session from a low of 9 to a high of 24. We translated counts to percentage scores for each session. In addition, observers also rated the *Student Engagement* dimension of the CLASS-S (see Quality of Delivery section for a description of the properties of the CLASS-S). Here, observers rated the degree that students were focused and attentive in the classroom and actively participating in each learning activity. In the low-range of this code, the majority of the students are disengaged from the class or distracted from learning. In the mid-range of this code, students appear to be passively engaged in the learning, not actively participating in the classroom; or there is a mix of student engagement in which some are engaged and others are not. In the high-range, the majority of the students are actively participating in the lesson.

2.5.6. Computing fidelity of receipt

Fidelity of receipt (α = .87) was obtained from the end-of-intervention student survey. We operationalized fidelity of receipt as student-reported confidence (1 = *not at all confident*, 5 = *very confident*) that they could engage in or demonstrate the skills highlighted in each session, and how much they agreed or disagreed (1 = *strongly disagree*, 5 = *strongly agree*) with the identity-based motivation messages regarding interpretation of difficulty and strategy development. The full set of items is provided in Table S2.15 in Supplemental Materials.

2.5.7. Computing Student-level and Classroom-level fidelity scores

Classroom-level (α = .85) fidelity is the mean of the five components of fidelity *dosage, adherence, student receptiveness, quality of delivery*, and *fidelity receipt,* with data coming from individual students for *quality of delivery* and *fidelity of receipt* averaged at the classroom-level. Student-level fidelity (α = .73) is the mean of the five components of fidelity *dosage, adherence, student receptiveness, quality of delivery*, and *fidelity receipt* with data from individual students for *quality of delivery* and *fidelity of receipt* maintained at the student-level.

2.6. Computing Core GPA and course failure

Chicago Public Schools provided student 7th and 8th grade grades for each marking period (n = 197 students had full records, 12 students were missing 7th-grade grades and an additional 2 students were missing 8th grade grades, likely due to out-of-district moves). We computed 7th -grade core GPA as an average of final grades for core classes (Math, Science, English, History, and Social Studies) in 7th-grade, and 8th-grade core GPA as an average of these core classes for 8th grade. We computed 7th grade course failure as 0 = no course failures in any marking period in 7th grade and 1 = at least one 7th-grade course failure in a marking period. We computed 8th-grade course failure as 0 = no course failures in any marking period in 8th grade and 1 = at least one 8th grade course failure in a marking period.

2.7. Computing teaching quality inside and outside the IBM intervention

We did not have the resources to video-record teachers in their subject classes outside the *Pathways* intervention so that a full direct comparison of general teaching quality and teaching quality within *Pathways* was not possible. However, in the end-of-intervention student survey we asked students to report on an element of teaching quality (*teacher-driven classroom climate*) for each of their subject teachers (math, science, English, and history). Students rated whether their teacher in each subject was enthusiastic, warm, clear, and knowledgeable (1 = strongly disagree, 5 = strongly agree, α = .85). To preserve independence of judgment, in our analyses, we compared ratings of each teacher from their *Pathways* students (describing them in *Pathways*) to the ratings each teacher received from their non-*Pathways* students in their subject class. The exact wording of each item is Table S2.16 in Supplemental Materials.

3. Results

3.1. Can teachers implement with fidelity?

On average, training appeared successful at attaining sufficient implementation fidelity: our brief two-day training resulted in average fidelity satisfying the 60% threshold criteria with some room for improvement, as detailed next. Overall, 89% of students experienced the intervention with at least 60% fidelity; the mean student-experienced fidelity was M = 68.71%, SD = 7.11. We present these results graphically in Fig. 2 by displaying the cumulative percentage of students at each level of fidelity. Results are consistent at the classroom level, 87.5% of classrooms (seven of eight) experienced fidelity above the 60% threshold and the eighth classroom had near threshold fidelity at 59%. For ease, classroom level results are presented in Table 3 from lowest to highest fidelity classroom, and in Fig. 3 as a boxplot. The boxes on the left are the classrooms and the final boxplot (colored gray) is the average across classrooms. In each boxplot, the top of the box shows the highest 25% fidelity, the bottom of the box shows the lowest 25%, and the dark line in the box shows median fidelity. The whiskers show the lowest and highest fidelity. As a can be seen, Fig. 3 graphically shows that fidelity across classrooms fits the Durlak and DuPre (2008) 60% threshold to 80% practical maximal range. On average, classrooms fidelity was 68.28% (SD = 5.95%). Student variation within classrooms
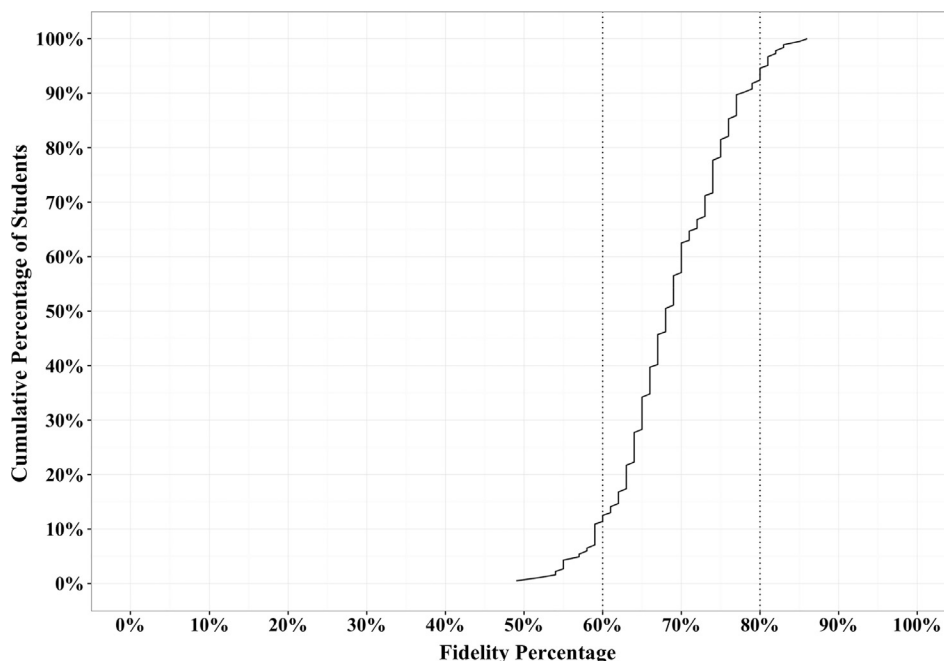
Fig. 2. Cumulative percentage of students by fidelity level.

Table 3
Classroom fidelity ordered from lowest fidelity classroom to highest fidelity classroom.

Classroom	Average fidelity	
	M	SD
1	59%	4%
2	64%	3%
3	66%	3%
4	67%	4%
5	69%	4%
6	69%	3%
7	76%	4%
8	77%	4%

is presented in Fig. 3 as whiskers representing the upper and lower limits of fidelity experienced by students in each classroom. Following Tukey (1977), whiskers exclude any extreme outliers that are beyond 1.5 times the size of the difference between the lowest 25% and the highest 75% above or below those quartiles; these outliers are represented individually as dots.

3.2. Distinguishing fidelity specific to Pathways

Before examining the effects of differences in *Pathways* fidelity on student outcomes, we addressed the question of whether teacher quality in *Pathways* was distinct from teacher quality in their subject matter classes. To do so, we used the only element of teacher quality outside of *Pathways* that we had, which was student ratings of *teacher-*
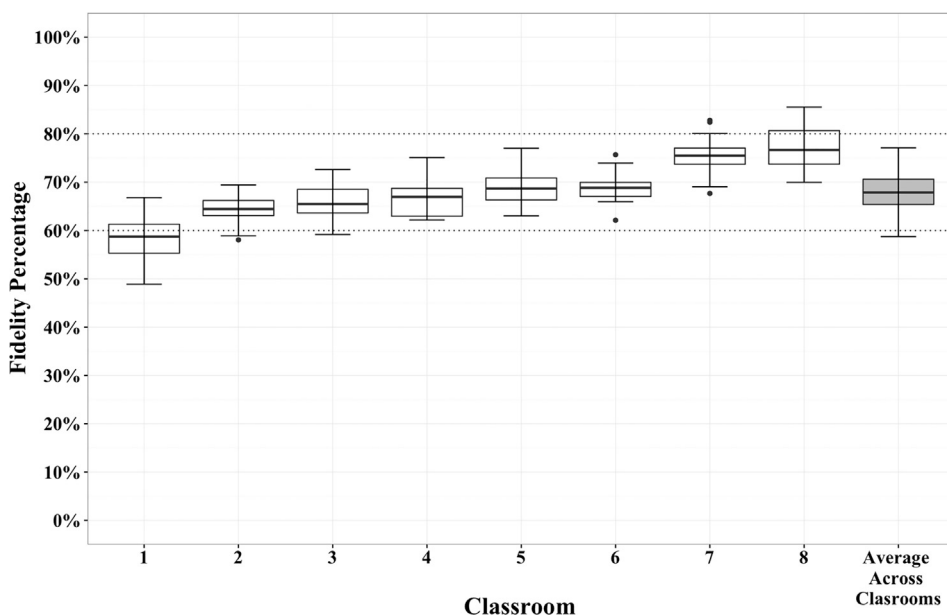

Fig. 3. Distribution of fidelity experienced by students in each classroom. *Note*: Classrooms are ordered from lowest to highest fidelity. The top and bottom of the box represent fidelity attained at the top 75% and bottom 25% respectively. The narrower the box, the more uniform the classroom experience is for students in this 25–75% range. The dark line inside the box highlights the median (middle 50%) attained classroom fidelity. The whiskers represent the full range of fidelity students experience in each classroom excluding outliers. We follow Tukey's (1977) definition of outlier scores. Outlier low scores are lower than the difference between the bottom 25% and 1.5× the difference between the lowest 25% and highest 75%. Outlier high scores are higher than the sum of the top 25% and 1.5× the difference between the lowest 25% and highest 75%. Outliers are plotted separately as individual dots (classroom 2, 6, and 7).

driven classroom climate in *Pathways* and in subject matter classes. To preserve independence of assessment, we used student ratings of their *Pathways* teacher as the teacher's *Pathways* quality rating. We used students who did not have that teacher for *Pathways* to obtain the teacher's subject class quality rating. Then we examined the correlation between the two ratings. We found a non-significant correlation $r = .15$, $p = .73$, which implies that teacher-level *Pathways* fidelity is not simply the product of a teacher's ability to create a positive classroom climate generally.

Further examination of the data revealed that half of teachers had higher classroom climate scores in *Pathways* than in their subject classroom and half had lower classroom climate scores in *Pathways* than in their subject classrooms (*Pathways* $M = 78.51$, $SD = 4.04$; subject $M = 80.53$, $SD = 5.43$). A meta-analytic synthesis using a random effects model, showed no overall significant pattern of differences, Cohen's $d = 0.16$, $SE = 0.15$, 95% CI (0.45, −0.14), $z = 1.023$, $p = .31$, as reflected in the nonsignificant difference and the fact that the 95% confidence interval includes zero. The implication is that teacher inside-of-*Pathways*-quality (fidelity) is distinct from teacher outside-of-*Pathways*-quality (as assessed by teacher-driven classroom climate). Hence our assessment of *Pathways* fidelity is not simply a reflection of a teacher trait or characteristic that is independent of training in *Pathways*. Having established that *Pathways* fidelity is unique, we now turn to the question of whether delivering *Pathways* with fidelity matters for student academic outcomes, as operationalized by core course grade point average and course failure rates.

3.3. Effects of fidelity

3.3.1. Preliminary analyses

We tested the effect of demographic variables on core course grade point average (Core GPA) and likelihood of course failure prior to testing the effects of fidelity on these variables. We did so in six regression equations testing different outcomes: (1) Core GPA at the end of 7th grade. (2) Core GPA at the end of 8th grade. (3) Core GPA at the end of 8th grade controlling for Core GPA at the end of 7th grade. (4) Any class failed in any marking period in 7th grade (1 = any failure, 0 = no failures). (5) Any class failed in any marking period in 8th grade (1 = any failure, 0 = no failures). (6) Any class failed in 8th grade controlling for any class failed in 7th grade. To test for effects of demographics we followed Cohen, Cohen, West, and Aiken (2003) and used contrast or effect codes in regression equations predicting Core GPA (regression equations 1 to 3) and dummy codes in logistic regression equations predicting course failure (regression equations 4 to 6). For regression equations, our contrast codes were (1 = female, −1 = male) and (free or reduced price lunch status 1 = eligible, −1 = not eligible). We created effect codes for each of the four racial-ethnic descriptors (Hispanic, Black, Asian, multiracial-ethnic) with White serving as the base group. So for example, the Hispanic effect code was Hispanic = 1, White = −1, Black = 0, Asian = 0, multiracial or multiethnic = 0. For logistic regression equations (1 = course failure, 0 = no course failure), our dummy codes were (1 = female, 0 = male), (1 = free or reduced price lunch eligible, 0 = not), (1 = identify as Hispanic, 0 = not), (1 = identify as African American, 0 = not), (1 = identify as Asian, 0 = not), and (1 = identify as multiracial or multi-ethnic, 0 = not). Each of these regression equations is presented in Section 3 of our Supplemental Materials. These regressions show that being female was associated with better outcomes and identifying as African American with worse outcomes. In addition, identifying as Hispanic was sometimes associated with better outcomes and receiving free or reduced lunch had a trend-level effect on 8th grade course failure. As a result, we report all analyses with these covariates included.

Table 4
Effects of student-level fidelity on 8th grade core GPA.

Predictor	B	SE	β	t	p	ΔR²	p
Two Step Model						.468	.00
Step 1							
7th Grade Final Core GPA	0.59	0.05	0.62	12.12	.00		
Imputed fidelity measures	−0.53	0.13	−0.22	−3.99	.00		
Imputed 7th Grade Core GPA	0.17	0.19	0.05	0.88	.38		
Step 2						.041	.00
Fidelity	0.02	0.01	0.21	4.14	.00		
Three Step Model							
Step 1						.468	.00
7th Grade Final Core GPA	0.59	0.05	0.62	12.12	.00		
Imputed fidelity measures	−0.53	0.13	−0.22	−3.99	.00		
Imputed 7th Grade Core GPA	0.17	0.19	0.05	0.88	.38		
Step 2						.140	.00
Hispanic	0.55	0.10	0.42	5.55	.00		
Black	−0.23	0.14	−0.11	−1.68	.10		
Asian	0.02	0.11	0.01	0.17	.86		
Multiracial-ethnic	−0.30	0.31	−0.10	−0.98	.33		
Free or reduced price lunch	0.12	0.08	0.07	1.47	.14		
Female	0.07	0.04	0.09	1.77	.08		
Step 3						.028	.00
Fidelity	0.02	0.01	0.17	3.90	.00		

Note: Fidelity is computed at the student-level.

3.3.2. Fidelity predicts core GPA

Fidelity predicted 8th-grade end-of-year Core GPA, whether or not demographic covariates were included. Specifically, each fidelity percentage increase is associated with an increased Core GPA of 0.02. Consider what would happen if fidelity increased from threshold level (60%) to practical maximum level (80%). This increase in fidelity would result in a .40 increase in Core GPA, the equivalent of moving from a C+ to almost a B. We used 2-Step (no demographic covariates) and 3-Step (with demographic covariates) hierarchical multiple regression analyses to test for effects of fidelity. In each regression equation, we first controlled for prior grades by entering at Step 1 student's final 7th-grade Core GPA and dummy codes for imputed data on fidelity and 7th-grade Core GPA. Then we asked if student-level fidelity mattered either by entering it at Step 2 in the 2-Step model or by entering it at Step 3, after first controlling for being female, free and reduced price lunch status, and identifying as Hispanic, as African American, as Asian, or as multiracial or multi-ethnic at Step 2.

Both models revealed that fidelity mattered for 8th-grade end-of-year Core GPA (Table 4, top panel, 2-Step model B = .024, SE = .006, β = .205, $p < .001$, 95% CI [.013, .036]; Table 4, bottom panel, 3-step model B = .020, SE = .005, β = .170, $p < .001$, 95% CI [.010, .030]). These effects remain significant and virtually unaltered if we exclude data from the small special education classroom or from students with imputed data, as detailed in Section 4 of the Supplemental Materials.

3.3.3. Does moving from threshold fidelity improve core GPA?

Our next analyses unpacked these positive effects of fidelity on Core GPA. We examined whether the significant effect of fidelity was due to the positive effect of fidelity for students and classrooms near the practical maximum of fidelity (80%) or if positive effects could already be seen at the mid-range between threshold and practical maximum. This is different from simply finding that higher fidelity matters since it pinpoints more specifically what level of fidelity training should target. As detailed next, we found that being near the practical maximum of fidelity mattered.
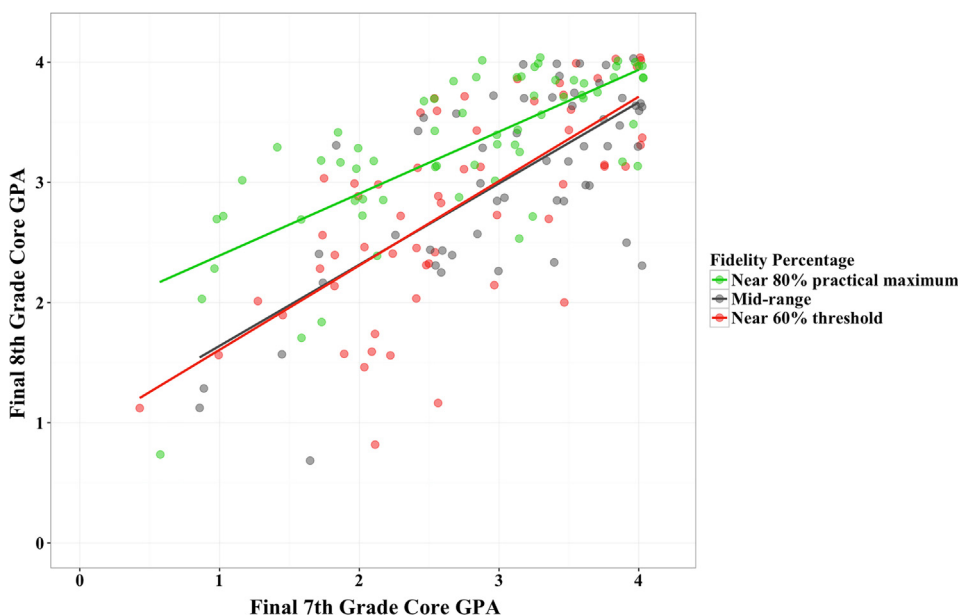
Fig. 4. The relationship between fidelity and Core GPA. *Note*: The green represents students receiving the intervention near practical maximum (top third, M = 76.31%), the red represents students receiving the intervention near threshold (bottom third, M = 61.08%), the gray represents students receiving the intervention in the mid-range (middle third, M = 68.10%).

We addressed this question by splitting our students into three equal groups based on their student-level fidelity scores. The bottom third (range 48.90–65.47%; $M = 61.33\%$, $SD = 3.72\%$) averaged at about what Durlak and DuPre (2008) described as threshold fidelity. The top third (range 70.86–85.54%; $M = 76.22\%$, $SD = 3.50\%$) averaged at about what Durlak and DuPre (2008) described as the 'practical maximum' of delivered fidelity. The middle third (65.47–70.80%; $M = 67.99\%$, $SD = 1.55\%$) averaged about midway between these. We used these three fidelity groups and ran two analyses of covariance (ANCOVA) models, one without demographic covariates ($F(2, 203) = 14.07$; $p < .001$, $\eta^2 = .12$), and one with demographic covariates ($F(2, 197) = 11.78$; $p < .001$, $\eta^2 = .11$). Each showed a significant effect of fidelity group on 8th grade end-of-year Core GPA, controlling for 7th grade core GPA and whether data were imputed (dummy variables). As detailed in Section 5 of the Supplemental Materials, these effects are significant and virtually unaltered if we exclude data from the small special education classroom or from students with imputed data.

We followed up with three planned contrasts, contrasting the practical maximum group to the lower threshold group and to the mid-range group, and contrasting the lower threshold group to the mid-range group. Given multiple comparisons, we applied Bonferroni adjustments to all *p*-values and confidence intervals. Being near the practical maximum mattered. The results of the planned contrast showed that the practical maximum group diverged from the other two groups. Being in the practical maximum fidelity group was associated with significantly higher Core GPA than being in the mid-range fidelity group. This result was found both for analyses without demographic covariates ($F(1, 203) = 16.73$, $p < .001$, 95% CI of the between-group difference [.164, .634], $\eta^2 = .08$) and for analyses with demographic covariates ($F(1, 197) = 8.98$, $p < .01$, 95% CI of the between-group difference [.052, .487], $\eta^2 = .04$). Being in the practical maximum fidelity group was also associated with significantly higher Core GPA than being in the near threshold fidelity group. This result was found both for analyses without demographic covariates ($F(1, 203) = 24.73$, $p < .001$, 95% CI of the between-group difference [.244, .703], $\eta^2 = .11$) and for analyses with demographic covariates ($F(1, 197) = 23.14$, $p < .001$, 95% CI [.202, .610], $\eta^2 = .11$). Core GPA did not differ for students in the near threshold fidelity group compared to students in the midrange fidelity group whether analyses were without demographic covariates $F(1, 203) = 0.60$, $p = 1.00$, 95% CI [−.158, .307], $\eta^2 = .00$ or with demographic covariates $F(1,197) = 2.49$, $p = .35$, 95% CI [−.073, .346], $\eta^2 = .01$.

For ease, we also represented these results graphically in Fig. 4, without demographic covariates and only for students with non-imputed data. We plotted students' 8th-grade end-of-year Core GPA as a function of their 7th-grade end-of-year Core GPA separately for each fidelity group. This allowed us to see the effect of fidelity over and above the effect of the prior year's academic outcome. Specifically, we plotted a dot for each student, with their 7th-grade Core GPA as the x-value and their 8th-grade Core GPA as the y-value. We used green colored dots for students who experienced near practical maximum fidelity ($M = 76.31\%$, $SD = 3.53\%$). The green regression line shows the predicted 8th-grade Core GPA given 7th-grade GPA for students experiencing near practical maximum fidelity. We used red colored dots for students who experienced near threshold fidelity ($M = 61.08\%$, $SD = 3.84\%$). The red regression line shows the predicted 8th-grade Core GPA given 7th-grade GPA for students experiencing near threshold fidelity. We used gray colored dots for fidelity at the mid-range between these two ($M = 68.10\%$, $SD = 1.48\%$). The gray regression line shows the predicted 8th-grade Core GPA given 7th-grade GPA for students experiencing mid-range fidelity. The effect of being in the near practical maximum group is easy to see by looking at the green colored regression line. As can be seen, the green line is above both the gray (mid-range group) and red (near lower threshold group) regression lines. The gray and red regression lines almost fully overlap. While visually, the green regression line is particularly divergent from the others for students who entered 8th grade with poorer 7th-grade core course grades, we do not find an interaction between prior grades and fidelity using the continuous measure of fidelity and non-imputed data (B = −.01, SE = .01, β = −.59, p = .22, 95% CI [−.018, .004]).

3.3.4. Fidelity predicts course failure

Fidelity predicted course failure. Specifically, each increase in a single percentage point of fidelity is associated with a 5.9% reduction in the odds of having even a single course failure (or alternatively, a 6.3% increase in the odds of passing every course in every marking period.) Consider what would happen if fidelity increased from threshold level (60%) to practical maximum level (80%). This increase in fidelity is associated with a reduction in the predictive probability of failing a course from about 28% to 10%. As detailed next, fidelity mattered whether or not demographic controls were used.

We used 2-Step (no demographic covariates) or 3-Step (with demographic covariates) logistic regression equations in these analyses. We used logistic regressions because course failure is a binary variable

Table 5
Effects of implementation fidelity on 8th grade course failure.

Predictor	B	SE	Exp(β)	Wald	p	ΔR²	p
Two Step Model							
Step 1						.138	.00
7th Grade Course Failure	1.68	0.34	5.35	23.92	.00		
Imputed fidelity measures	0.80	0.49	2.23	2.70	.10		
Imputed 7th Grade Course Failure	−0.15	0.70	0.86	0.05	.83		
Step 2						.023	.02
Fidelity	−0.06	0.03	0.94	5.46	.02		
Three Step Model							
Step 1						.138	.00
7th Grade Course Failure	1.68	0.34	5.35	23.92	.00		
Imputed fidelity measures	0.80	0.49	2.23	2.70	.10		
Imputed 7th Grade Course Failure	−0.15	0.70	0.86	0.05	.82		
Step 2						.056	.03
Hispanic	0.32	0.76	1.38	0.18	.67		
Black	1.44	0.90	4.21	2.56	.11		
Asian	−0.20	0.88	0.82	0.05	.82		
Multiracial-ethnic	1.51	1.66	4.53	0.83	.36		
Free or reduced price lunch	−1.39	0.77	0.25	3.29	.07		
Female	−0.77	0.36	0.46	4.55	.03		
Step 3						.022	.02
Fidelity	−0.06	−.03	0.94	5.49	.02		

Note: Fidelity is computed at the student-level; ΔR² = change in Cox & Snell R-squared.

(1 = course failure, 0 = no course failure). In each equation, we first controlled for prior course failure by entering at Step 1 a dummy variable for a student's failure in any course in any marking period in 7th grade and dummy codes for imputed data on fidelity and 7th grade Core GPA. Then we asked if student-level fidelity mattered either by entering it at Step 2 in the 2-Step model or by entering it at Step 3 after first controlling for being female, free and reduced price lunch status, and identifying as Hispanic, as African American, as Asian, or as multiracial or multi-ethnic at Step 2. Fidelity predicted course failure to the same extent in both models: no demographic controls model B = −.058, SE = .025, Wald = 5.46, Exp(B) = .943, p < .02, 95% CI [.898, .991] and demographic controls model B = −.061, SE = .026, Wald = 5.49, Exp(B) = .941, p < .02, 95% CI [.894, .990]). Table 5 provides the details of both models (the 2-Step model, top panel; the 3-Step model, bottom panel). Effects remain significant and virtually unaltered in size when analyses exclude the smaller special education classroom or exclude students with imputed data as detailed in Section 6 of Supplemental Materials.

3.3.5. Does moving from threshold fidelity improve course failure rates?

We examined whether the significant effect of fidelity on course failure was stepwise, due to fidelity's positive effect when it was near the 80% practical maximum or if fidelity's positive effects were more linear and could already be seen at mid-range fidelity. Course failure is a binary construct (fail, pass) so we used logistic regression equations, creating dummy codes to represent the three fidelity groups. We tested two comparisons – the effect of being in the near-practical maximum group or in the midrange group relative to the near-threshold group and the effect of being in the near-threshold group relative to the midrange group. To test for stability of effects with and without demographic controls, we ran each comparison twice, once as a 2-step model (without demographic controls) and again as a 3-step model (including demographic controls). As detailed in Tables 6 and 7, we found weak evidence for a stepwise effect of fidelity for risk of course failure, implying that the relationship between fidelity and risk of course failure may be more linear.

Table 6
Effects of implementation fidelity group on 8th grade course failure, compared to near-threshold group.

Predictor	B	SE	Exp(β)	Wald	p	ΔR²	p
Two step model							
Step 1						.138	.00
7th Grade course failure	1.68	0.34	5.35	23.92	.00		
Imputed fidelity measures	0.80	0.49	2.23	2.70	.10		
Imputed 7th grade course failure	-0.15	0.70	0.86	0.05	.83		
Step 2						.030	.02
Midrange group	-0.44	0.41	0.65	1.14	.29		
Near-practical maximum group	-1.14	0.43	0.32	7.07	.01		
Three step model							
Step 1						.138	.00
7th Grade course failure	1.68	0.34	5.35	23.92	.00		
Imputed fidelity measures	0.80	0.49	2.23	2.70	.10		
Imputed 7th grade course failure	-0.15	0.70	0.86	0.05	.82		
Step 2						.056	.03
Hispanic	0.32	0.76	1.38	0.18	.67		
Black	1.44	0.90	4.21	2.56	.11		
Asian	-0.20	0.88	0.82	0.05	.82		
Multiracial-ethnic	1.51	1.66	4.53	0.83	.36		
Free or reduced price lunch	-1.39	0.77	0.25	3.29	.07		
Female	-0.77	0.36	0.46	4.55	.03		
Step 3						.032	.01
Midrange group	-0.37	0.44	0.69	.729	.39		
Near-practical maximum group	-1.27	0.45	0.28	7.92	.01		

Note: Fidelity is computed at the student-level; ΔR² = change in Cox & Snell R-squared.

As detailed in Table 6, students in the near-practical maximum group were less likely to fail a course than those in the near-threshold group: model without demographics, B = −1.14, SE = .429, Wald = 7.07, Exp(B) = .319, p < .01, 95% CI [.138, .741]; model with demographic controls, B = −1.266, SE = .450, Wald = 7.92, Exp (B) = .292, p < .01, 95% CI [.117, .681]. However, likelihood of course failure did not differ between students in midrange and the near-threshold groups: model without demographics, B = −.435, SE = .407, Wald = 1.140, Exp(B) = .647, p = .29, 95% CI [.292, 1.438]; model with demographic controls, B = −.374, SE = .438, Wald = .729, Exp

Table 7
Effects of implementation fidelity group on 8th grade course failure, compared to midrange group.

Predictor	B	SE	Exp(β)	Wald	p	ΔR²	p
Two Step Model							
Step 1						.138	.00
7th Grade course failure	1.68	0.34	5.35	23.92	.00		
Imputed fidelity measures	0.80	0.49	2.23	2.70	.10		
Imputed 7th grade course failure	-0.15	0.70	0.86	0.05	.83		
Step 2						.030	.02
Near-threshold group	0.44	0.41	1.54	1.14	.29		
Near-practical maximum group	-0.71	0.46	0.49	2.34	.13		
Three step model							
Step 1						.138	.00
7th Grade course failure	1.68	0.34	5.35	23.92	.00		
Imputed fidelity measures	0.80	0.49	2.23	2.70	.10		
Imputed 7th grade course failure	-0.15	0.70	0.86	0.05	.82		
Step 2						.056	.03
Hispanic	0.32	0.76	1.38	0.18	.67		
Black	1.44	0.90	4.21	2.56	.11		
Asian	-0.20	0.88	0.82	0.05	.82		
Multiracial-ethnic	1.51	1.66	4.53	0.83	.36		
Free or reduced price lunch	-1.39	0.77	0.25	3.29	.07		
Female	-0.77	0.36	0.46	4.55	.03		
Step 3						.032	.01
Near-threshold group	0.37	0.44	1.453	.729	.39		
Near-practical maximum group	-0.89	0.51	0.41	3.10	.08		

Note: Fidelity is computed at the student-level; ΔR² = change in Cox & Snell R-squared.

(B) = .688, p = .39, 95% CI [.292, 1.623]). Similarly, as detailed in Table 7, likelihood of course failure did not differ between students in midrange and near-practical maximum groups: model without demographics, B = −0.71, SE = .462, Wald = 2.34, Exp(B) = .493, p = .13, 95% CI [.199, 1.219]; model with demographic controls, B = −0.892, SE = .506, Wald = 3.10, Exp(B) = .410, $p < .08$, 95% CI [.152, 1.106]). Effects are virtually unaltered if we exclude data from the small special education classroom (analyses presented in Section 7 of the Supplemental Materials). Although interpretability of these results is limited due to small sample size, these findings imply that the effect of fidelity on course failure is linear, rather than clearly stepwise. As detailed in the supplemental materials, one set of analyses does show a more pronounced stepwise effect; the effect of being in the near-practical maximum group compared to being in the midrange group is significant in the two step model (excluding demographic controls) when students with imputed data are excluded from analyses (B = −1.264, SE = .543, Wald = 5.411, Exp(B) = .283, p = .02, 95% CI [.097, 1.219]).

3.4. Teachers' perspectives

AIR staff met with each teacher separately to discuss usability (how easy it was to use the program given the provided training, resources, and support) and feasibility (how well teachers felt they could implement the program given other demands in their teaching context) using a two-page set of structured open-ended probes (see, Section 1, Supplemental Materials). The first question asked teachers what they liked about *Pathways* before shifting to questions about usability and feasibility that targeted areas for improvement. In response to the first question, teachers said that they liked, loved, or enjoyed it, that it was well-designed and well thought out, that students were disappointed when it ended, that "It was a good platform for the students to start to see a structure to get them to look at where they are going in the future", and that it "opened minds and eyes to next steps." When asked how to improve it, they were unanimous in three ideas as to how to increase their fidelity of implementation. First, they suggested that the intervention manual itself be reformatted to look like other teacher materials so that it would be easier for them to process the information. Second, that the materials for students should be reusable (e.g., laminated worksheets rather than single use). Third, that they should be provided PowerPoint rather than newsprint for each session and activity. Finally, though not mentioned by each of the teachers, a number also suggested changing the training to three days to provide teachers more time to practice and absorb the intervention. When asked to detail problems needing improvement, teachers also gave a variety of idiosyncratic critiques—critiques that were unique to a single teacher. Unlike the unanimity of the other responses, this variability led us to consider whether what teachers said was related to their fidelity overall or their fidelity in any particular session. We did not find a clear pattern; it was not that teachers singled out sessions they delivered with lower fidelity or that teachers all had problems with the same sessions. Hence, we also viewed videotape to understand sessions in which delivery was problematic and consider ways to improve the manual and training to address these issues.

4. Discussion

We found that teachers can deliver identity-based motivation intervention with fidelity in their classrooms and that higher fidelity matters. Fidelity changes students' academic trajectories particularly when it is nearer the 'practical maximum' of 80% found by Durlak and DuPre's (2008) examination of meta-analyses of practitioner-delivered interventions. Our findings suggest that feasible increases in fidelity (moving fidelity closer to 80%) have meaningful effects on core grade point average and course failure rates. Teachers had consensus suggestions for improving fidelity that we implemented for future use.

Moreover, after viewing videorecorded classroom sessions, we could discern what future training effort should focus on. Our findings are important for a number of reasons. First, harnessing students' identity-based motivation matters for academic outcomes. Second, our findings are necessary first steps for embarking on future 'gold standard' randomized control tests of the effect of identity-based motivation intervention. Third, our findings highlight the need to assess fidelity in ways that allow unpacking it to understand how to improve fidelity in the future.

4.1. Summary of results

We asked if a two-day abbreviated version of the five-day training used for non-teachers was sufficient for teachers to attain threshold fidelity and if higher attained fidelity changed students' academic trajectories. Based on the literature, we operationalized fidelity as five interdependent components of *dosage*, *adherence*, *quality of delivery*, *student responsiveness*, and *fidelity-of-receipt*. When teachers deliver the planned number of sessions when planned (*dosage*) and with the correct content (*adherence*) in the correct way (*quality of delivery*), their students should respond with attention, engagement, and productivity (*student responsiveness*) and hence internalize the take-home points (*fidelity of receipt*). We carefully assessed fidelity with reliable, structured measures coded from videotape of each session (observer report) and student report. To do so, we used the original fidelity instruments used in the trainer-led *School-to-Jobs* and incorporated the CLASS-S given developments in the field of education and its high overlap with the original instruments.

We followed other educational intervention evaluations by operationalizing fidelity as sufficient if it met Durlak and DuPre's (2008) empirically derived threshold of 60% (e.g. Bloomquist et al., 2013; Lindsay et al., 2017). Given that Durlak and DuPre (2008) also found that non-researchers rarely deliver with fidelity above 80%, making 80% a practical maximum of fidelity, we asked if getting closer to this practical maximum mattered. We found that our two-day training was successful: Almost all students received the *Pathways* identity-based motivation intervention with fidelity at or above threshold and average classroom-level fidelity was within the Durlak and DuPre suggested range of 60% to 80%. Moreover, higher fidelity mattered. We examined the differential effects of near threshold, near practical maximum, and mid-range fidelity. Though Durlak and DuPre noted that higher fidelity matters, their analyses are general and do not separately examine if moving from threshold (60%) to practical maximum (80%) fidelity has an effect on intended outcomes. Our review of the literature since then did not uncover anyone else examining this possibility. For core grade point average, own results suggest that moving from threshold to practical maximum fidelity does matter for academic performance. Students who received *Pathways* at close to practical maximum fidelity had better academic outcomes than those who did not. For risk of course failure, our results suggest a more linear and less stepwise effect of increased fidelity on reduced risk of course failure. These analyses showed improved end-of-year 8th grade core grade point average and reduced likelihood of course failure, controlling for grades and course failure in 7th grade.

Results were robust to inclusion of demographic controls of gender, race, and free or reduced price lunch and to inclusion or exclusion of special education classroom or students with imputed data. Effects were found for students at every level of 7th-grade grade point average. Visually, effects looked stronger for students with lower 7th-grade core course grade point averages, but we did not find an interaction between fidelity and prior academic performance in our sample.

Our end of *Pathways* intervention feedback from teachers was positive. They liked *Pathways* and found it usable, and had very useful and actionable suggestions to improve, which we detail after considering theoretical implications of our results. Because we had a video record of each session, we could closely examine quality of delivery and consider

ways to improve training to more fully engage teachers with core aspects of identity-based motivation theory (e.g. interpretation of experienced difficulty as importance).

4.2. Theoretical implications: identity-based motivation

Our results add to the literature on the importance and malleability of identity-based motivation. Because identities are experienced as stable but are in fact dynamically constructed in context, small contextual cues can trigger important changes. For example, student's next year and adult possible selves can be made to feel near and connected to what they are doing right now rather than far away and irrelevant to right now (e.g. Nurra & Oyserman, 2018). Similarly, students can be cued to use a difficulty-as-importance mindset in making sense of experiences of difficulty with schoolwork and in considering whether school-focused possible identities are really 'for me' or 'for us' (Oyserman et al., 2018; Smith & Oyserman, 2015). When led to consider their future selves as relevant to right now and to interpret experienced difficulty as a sign that these future selves are important and that failures along the way are normal, students succeed.

Without intervention, students may experience their future selves as far and irrelevant to right now, and may misinterpret difficulties along the way as implying that school-focused identities are not for them (Oyserman, 2015a, 2015b). Even one-time cues such as those used in experiments can be powerful, influencing these elements of identity-based motivation, and through influencing elements of IBM, changing student focus and effort, and impacting grade point average few months later. Of course, the kinds of one-time cues used in experiments are not enough for effects to last over years. For that, intervention is needed so that students are repeatedly exposed to cues, shaping their focus, and hence how they likely will make sense of their experiences over time. Prior intervention research revealed that university students and adults with undergraduate degrees can successfully turn on middle schoolers' identity-based motivational processes with a brief manualized identity-based motivation intervention (Oyserman et al., 2006; Oyserman et al., 2002). Our results extend these findings by demonstrating that classroom teachers can implement identity-based motivation as part of the regular school day.

4.3. Practical implications

Our fidelity analyses allowed us to move beyond documenting that we could attain threshold levels of fidelity to more carefully unpack whether there is a benefit of moving beyond threshold fidelity, and if there is, at what level this benefit accrues. We found that attaining higher fidelity matters for students' academic trajectories, and that the effect of higher fidelity was concentrated at fidelity closer to the practical maximum of 80%. There are a number of important practical implications of these results. First, for identity-based motivation researchers, our results imply that it is worth investing in teacher professional development to support increased fidelity of receipt and of delivery of identity-based motivation. Second, our careful assessment of fidelity and our separate request for teacher feedback worked synergistically to allow us to respond to both teacher-noted and researcher-noted points for improvement. Using both methods also highlighted the limits and strengths of each source of information. Teachers can notice and report on what feels difficult but cannot know why something felt difficult. It could be that their preparation was insufficient or that the training was insufficient or that the session itself was problematic. The same issue arises when something feels easy – the teachers can notice that but cannot know if that is because the session went well or because whatever they did was fun and easy. Just because things feel fine does not necessarily mean that the intervention is delivered or received as intended (e.g., the so-called Dunning-Kruger effect, Kruger & Dunning, 1999). Researchers can notice and report on variability in fidelity, highlighting when it is higher or lower, but not

know if teachers experienced difficulty or ease at these points. Improving training requires both teacher and researcher perspectives.

After implementing *Pathways*, our teachers were encouraged to highlight any problems they had with sessions. All reported that they liked *Pathways*, found it worth the time to get trained, and also suggested practical ways to improve usability and feasibility. Teachers made pragmatic suggestions to improve fidelity that unanimously focused on creating a teacher's implementation manual and delivery system (e.g. pre-prepared Powerpoint rather than pre-prepared newsprint) that felt more similar to their current textbooks. Teachers also articulated what felt difficult and which session activities did not work well for them. These points differed by teacher. Both unanimous and teacher-specific points were valuable to us because teacher suggestions were different from the suggestions we as the research team had for improving fidelity after examining the videos. Without teacher feedback we would not have known to change the manual or to switch to PowerPoint and reusable materials. Without the video records we would not have been able to understand where exactly training should be improved because teachers, like all people, do not know what they do not know, and this is particularly true for novices who have low expertise (the so-called Dunning-Kruger effect; Dunning, 2011; Kruger & Dunning, 1999). Teachers can articulate what feels difficult but cannot be expected to know if the source of experienced problems is in their training, in their preparation, in their delivery, or in the session or activity itself. The same holds for ease: teachers can report what felt easy but not if that was because they delivered the activity as intended and everything worked. Hence the research team watched the video records of all sessions to learn where there were gaps in the training and where delivery fell short, separate from teacher-reported ease or difficulty.

Third, our results suggest that brief training can yield fidelity separate from characteristics of teachers. We base this implication on lack of association between *Pathways* fidelity and teacher core subject and lack of association between *Pathways* fidelity and teaching quality outside of *Pathways*. Our teachers taught each of the core subjects and we did not see a by-subject difference in fidelity. In addition, our analyses of the relationship between teacher quality ratings in their subject class and in *Pathways* revealed that teacher's quality ratings in their subject classes did not predict their quality ratings in *Pathways*.

Fourth, our results might generalize to other educational intervention evaluation efforts. Our review of the literature did not uncover other evaluation studies that unpacked the association of fidelity with outcomes by examining whether moving from threshold (60%) to practical maximum (80%) fidelity mattered for intended outcomes. This is a more sensitive analyses than simply documenting that more fidelity is better than less fidelity because it takes as a starting point that fidelity lower than 60% is insufficient and spotlights whether fidelity above 60% yields better effects than fidelity at 60%. It is unclear whether prior analyses address this issue because segmented analyses were not presented. An implication of our finding is that evaluators should ask if increasing fidelity above threshold has a linear effect on outcomes. If it does, then a careful examination of how to improve fidelity to move it beyond 60% threshold and closer to 80% practical maximum is warranted. If it does not, then there is no need to invest resources to improve fidelity beyond threshold.

Fifth, our results also have implications for the intervention fidelity literature. In our review of the literature we found both agreement as to what fidelity entails and diversity as to how fidelity is measured (e.g., O'Donnell, 2008; Durlak & DuPre, 2008). We documented that each of the five components of fidelity (*dosage*, *adherence*, quality of delivery, *student responsiveness*, and *fidelity of receipt*) can be clearly operationalized and reliably assessed with multiple measures. We benchmarked our five-component operationalization of fidelity against the empirically derived 60% threshold and 80% practical maximum of fidelity found by Durlak and DuPre (2008). We present empirical support both for the 60% to 80% range and document that higher fidelity, closer

to the practical maximum matters.

Finally, our results have a number of practical implications for feasible and scalable teacher training. To be feasible, training must be brief. To be effective, teachers need opportunities to practice, need to receive specific feedback while training, and need to have ongoing support while implementing (Darling-Hammond, Hyler, & Gardner, 2017). To be scalable, training should not be limited to a few trainers. To be useful, training should feel relevant to teacher practice. In the case of *Pathways,* teachers had some opportunity to practice as part of the brief 2-day training and as part of the weekly call-ins, but an additional training day focused on implementation and structured feedback might enhance teacher fidelity and indeed, that is what some teachers asked for. Some teachers carried *Pathways* terms and concepts into their classes throughout the remainder of the academic year, suggesting that *Pathways* provided a new way for teachers to engage their students about connecting school to their futures. To support scalability, the possibility of a teacher-trained as well as a teacher-led *Pathways* needs to be tested. We used our teacher feedback and examination of videotape to develop a web-based resource including preparation tips from teachers who delivered with fidelity, teacher-viewable videotape of high fidelity delivery, and a video-assisted structured training module that teachers who already delivered *Pathways* can use to train other teachers.

4.4. Limitations

As with any study, our study has a number of limitations. First, our analyses include eight classrooms and about two hundred students. While not small by standards of psychological research, replication of our results is necessary since any one study alone cannot provide a fully stable estimate of effects. Hence replication of our basic test is important and a goal that we are currently pursuing in our ongoing work. Second, our sample size meant we were not powered for mediation analyses, or to test for possible moderation of the effect of the intervention for previously lower versus higher performing students. Instead, as noted, our goal was to test the prediction that we could train teachers to attain threshold fidelity and so our study design focused on training. In doing so we addressed a gap in the literature on interventions, which is that fidelity is often not addressed at all (e.g., Chao, Visaria, Mukhopadhyay, & Dehejia, 2017) or is not assessed carefully enough to provide an empirical assessment of how much of an intended intervention students received (e.g., de Jong, Jellesma, Koomen, & de Jong, 2016; Bradley, Crawford, & Dahill-Brown, 2016). Our design highlights fidelity assessment, and addresses the question of whether threshold fidelity is likely to be attained with brief training. This is a distinct question from the typical evaluation research question, which ignores fidelity and focuses on student outcomes: whether students randomized to an IBM intervention group outperform students randomized to a no-IBM control group. We could not test this latter question in our fidelity-focused design since all students received intervention and therefore we did not randomize to experimental and control groups.

Separate from limitations to our sample and design, there were also a number of limitations to our training. We did not randomize teachers to varying intensity of training; instead we chose as our start-point what we thought was the minimal training likely to provide teachers with sufficient chances to learn. Teachers were provided a very brief, two-day training at their school. This 14-hour training truncated the 5-day, 40-hour training that trainers received in the *Schools-to-Job* intervention. While we were able to show that we could attain fidelity of at least threshold level, we did not test what would have happened with longer training. While our school-based method would likely be feasible were training to become teacher-led, we did not test what would have happened if teachers came together across schools or received longer training. These changes might have increased *quality of delivery* by exposing teachers to more diversity of styles and giving them more chance

to practice. Lastly, we provided a single training and did not assign teachers to different combinations of training for *adherence* vs. *quality of delivery*. This means that our study cannot shed light on how the components of fidelity might interact with each other.

4.5. Future directions

Our current results suggest three important future directions for research: randomized control test of outcomes, test of mediation, and development of new platforms for intervention scaling. A randomized control test would allow us to know whether students' academic trajectories change as a result of being randomly assigned to *Pathways* compared to school as usual or to an alternative socio-emotional or motivational intervention. A mediation test would allow us to know whether effects are due to changes in the three components of identity-based motivation (dynamic construction, action-readiness, and procedural-readiness). That is, whether changes are due to changes in the extent that students experience their future selves as connected to the present via schoolwork (dynamic construction). Whether changes are due to the extent that students take action to start and persist in their schoolwork (action-readiness). And finally, whether changes are due to the extent that students are flexibly able to interpret their experiences of difficulty as signals of importance rather than impossiblity (procedural-readiness). Another future direction is to test whether fidelity can be maintained when training is teacher-led rather than researcher-led as it was in the current iteration. A teacher-led training paradigm is clearly more usable and feasible for scaling as long as it yields adequate fidelity, something that future research should test.

5. Conclusion

Our results shed light on both the promise of scalability and the difficulty of scaling promising tests of theory in schools. We show that an identity-based motivation intervention can be delivered and experienced at above threshold fidelity after a feasibly brief 2-day intervention. We also show that moving from a threshold level of fidelity to higher fidelity matters. Specifically, an average shift from threshold to practical maximum fidelity is associated with a shift translating from a C + to almost a B core course grade point average and reduction in the predicted probability of failing a class from about 28% to 10%. Students across the continuum from high attaining through those with individualized educational programs benefit from the intervention. The implication is that teachers can help students harness their own high aspirations using identity-based motivation. When teachers help students imagine school as the path to their future, conceptualize strategies to succeed on that path, and see obstacles and failures along the way as signaling importance and value, they are helping students succeed academically. Given the meaningful size of effects, future work on scaling is critical.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.cedpsych.2018.04.004.

References

Aelenei, C., Lewis, N. A., & Oyserman, D. (2017). No pain no gain? Social demographic correlates and identity consequences of interpreting experienced difficulty as importance. *Contemporary Educational Psychology, 48,* 43–55.

Allen, J.P, Gregory, A., Mikami, A., Lun, J., Hamre, B., & and Pianta, R.C. (n.d.) Predicting Adolescent Achievement with the CLASS-S Observation Tool. Center for Advanced Study of Teaching and Learning, University of Virginia Curry School of Education. Retrieved from http://curry.virginia.edu/research/centers/castl/projects/castl-research-briefs.

Allensworth, E. M., & Easton, J. Q. (2005). *The on-track indicator as a predictor of high school graduation.* Chicago, IL: Consortium on Chicago School Research, University of Chicago Retrieved from http://consortium.uchicago.edu/sites/default/files/

publications/p78.pdf.

Allensworth, E. M., & Easton, J. Q. (2007). *What matters for staying on track and graduating in Chicago public high schools*. Chicago, IL: Consortium on Chicago School Research Retrieved from http://files.eric.ed.gov/fulltext/ED498350.pdf.

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., ... Czajkowski, S. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology, 23*(5), 443.

Bloomquist, M. L., August, G. J., Lee, S. S., Lee, C. Y. S., Realmuto, G. M., & Klimes-Dougan, B. (2013). Going-to-scale with the Early Risers conduct problems prevention program: Use of a comprehensive implementation support (CIS) system to optimize fidelity, participation and child outcomes. *Evaluation and Program Planning, 38*, 19–27.

Botvin, G. J., Baker, E., Dusenbury, L., Botvin, E. M., & Diaz, T. (1995). Long-term follow-up results of a randomized drug abuse prevention trial in a white middle-class population. *Journal of the American Medical Association, 273*(14), 1106–1112.

Botvin, G. J., & Griffin, K. W. (2015). Preventing tobacco, alcohol, and drug abuse through Life Skills Training. In L. M. Scheier (Ed.). *Handbook of drug abuse prevention research, intervention strategies, and practice*. Washington DC: American Psychological Association.

Bradley, D. N., Crawford, E. P., & Dahill-Brown, S. E. (2016). Defining and assessing FoI in a large-scale randomized trial: Core components of values affirmation. *Studies in Educational Evaluation, 49*, 51–65.

Brehm, S. S., & Brehm, J. W. (2013). *Psychological reactance: A theory of freedom and control*. New York, NY: Academic Press.

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation, 31*(2), 199–218.

Chao, M. M., Visaria, S., Mukhopadhyay, A., & Dehejia, R. (2017). Do rewards reinforce the growth mindset?: Joint effects of the growth mindset and incentive schemes in a field intervention. *Journal of experimental psychology: General, 146*(10), 1402–1419.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed). Mahwah, NJ: Lawrence Erlbaum Associates.

Crosse, S., Williams, B., Hagen, C. A., Harmon, M., Ristow, L., DiGaetano, R., ... & Derzon, J. H. (2011). Prevalence and Implementation Fidelity of Research-Based Prevention Programs in Public Schools. Final Report. Washington, D.C.: Office of Planning, Evaluation and Policy Development, U.S. Department of Education. Retrieved From http://files.eric.ed.gov/fulltext/ED529062.pdf.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*(1), 23–45.

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective Teacher Professional Development*. Palo Alto, CA: Learning Policy Institute Retrieved from https://learningpolicyinstitute.org/sites/default/files/product-files/Effective_Teacher_Professional_Development_REPORT.pdf.

de Jong, E. M., Jellesma, F. C., Koomen, H. M., & de Jong, P. F. (2016). A values-affirmation intervention does not benefit negatively stereotyped immigrant students in the Netherlands. *Frontiers in Psychology, 7*.

Destin, M. (2017). An open path to the future: Perceived financial resources and school motivation. *The Journal of Early Adolescence, 37*, 1004–1031.

Destin, M., & Oyserman, D. (2009). From assets to school outcomes: How finances shape children's perceived possibilities and intentions. *Psychological Science, 20*(4), 414–418.

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. *Advances in Experimental Social Psychology, 44*, 247–296.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*(3–4), 327–350.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research, 18*(2), 237–256.

Elmore, K. C., & Oyserman, D. (2012). If 'we' can succeed, 'I' can too: Identity-based motivation and gender in the classroom. *Contemporary Educational Psychology, 37*(3), 176–185.

Elmore, K., Oyserman, D., Smith, G. C., & Novin, S. (2016). When the going gets tough: Implication of reactance for interpretations of experienced difficulty in the classroom. *AERA Open, 2*(3), 1–11.

Fagan, A. A., Hanson, K., Hawkins, J. D., & Arthur, M. W. (2009). Translational research in action: Implementation of the Communities That Care prevention system in 12 communities. *Journal of Community Psychology, 37*(7), 809–829.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*(3), 613–619.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of

fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness, 2*(1), 88–110.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Landau, M. J., Oyserman, D., Keefer, L. A., & Smith, G. C. (2014). The college journey and academic engagement. *Journal of Personality and Social Psychology, 106*(5), 679–698.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lindsay, J., Davis, E., Stephan, J., & Proger, A. (2017). Impacts of Ramp-Up to Readiness after one year of implementation (REL 2017–241). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. Retrieved from http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=1461.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: development, measurement, and validation. *Am. J. Eval. 24*(3), 315–340.

Nurra, C., & Oyserman, D. (2018). From future self to current action: An identity-based motivation perspective. *Self & Identity, 17*(3), 343–364.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Rev. Edu. Res. 78*(1), 33–84.

Oyserman, D. (2007). Social identity and self-regulation. In A. Kruglanski, & T. Higgins (Eds.). *Handbook of Social Psychology* (pp. 432–453). (2nd ed.). NY: Guilford Press.

Oyserman, D. (2015a). Identity-based motivation. In R. Scott, & S. Kosslyn (Eds.). *Emerging Trends in the Social Sciences*. Hoboken, NJ: John Wiley and Sons.

Oyserman, D. (2015b). *Pathways to success through identity-based motivation.* NY, New York, USA: Oxford University Press.

Oyserman, D., Bybee, D., & Terry, K. (2006). Possible selves and academic outcomes: How and when possible selves impel action. *Journal of Personality and Social Psychology, 91*(1), 188–204.

Oyserman, D., & Destin, M. (2010). Identity-based motivation: Implications for intervention. *The Counseling Psychologist, 38*(7), 1001–1043.

Oyserman, D., Destin, M., & Novin, S. (2015). The context-sensitive future self: Possible selves motivate in context, not otherwise. *Self and Identity, 14*(2), 173–188.

Oyserman, D., Elmore, K., Novin, S., Fisher, O., & Smith, G. C. (2018). Guiding people to interpret their experienced difficulty as importance highlights their academic possibilities and improves their academic performance. *Frontiers in Psychology, 9* Article 781.

Oyserman, D., Johnson, E., & James, L. (2011). Seeing the destination but not the path: Effects of socioeconomic disadvantage on school-focused possible self content and linked behavioral strategies. *Self and Identity, 10*(4), 474–492.

Oyserman, D., & Lewis, N. A. (2017). Seeing the destination AND the path: Using identity-based motivation to understand and reduce racial disparities in academic achievement. *Social Issues and Policy Review, 11*(1), 159–194.

Oyserman, D., Lewis, N. A., Jr, Yan, V. X., Fisher, O., O'Donnell, S. C., & Horowitz, E. (2017). An identity-based motivation framework for self-regulation. *Psychological Inquiry, 28*(2–3), 139–147.

Oyserman, D., Terry, K., & Bybee, D. (2002). A possible selves intervention to enhance school involvement. *Journal of Adolescence, 25*(3), 313–326.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Vol. Ed.), *Advances in Experimental Social Psychology: Vol. 19*, (pp. 123–205). New York: Academic Press.

Pianta, R. C., Hamre, B., Hayes, N., Mintz, S., & LaParo, K. M. (2008). *Classroom Assessment Scoring System - Secondary (CLASS-S)*. University of Virginia.

Preschool Curriculum Evaluation Research Consortium (2008). Effects of Preschool Curriculum Programs on School Readiness (NCER 2008-2009). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Resnick, B., Bellg, A. J., Borrelli, B., De Francesco, C., Breger, R., Hecht, J., ... Ogedegbe, G. (2005). Examples of implementation and evaluation of treatment fidelity in the BCC studies: Where we are and where we need to go. *Annals of Behavioral Medicine, 29*(2), 46–54.

Riordan, J., Lacireno-Paquet, N., Shakman, K., Bocala, C., & Chang, Q. (2015). Redesigningteacher evaluation: Lessons from a pilot implementation (REL 2015–030). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. Retrieved from http://ies.ed.gov/ncee/edlabs.

Smith, G. C., & Oyserman, D. (2015). Just not worth my time? Experienced difficulty and time investment. *Social Cognition, 33*(2), 1–18.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Velasco, V., Griffin, K. W., Botvin, G. J., Celata, C., & Lombardia, G. L. (2017). Preventing adolescent substance use through an evidence-based program: Effects of the Italian adaptation of Adaptation of Life Skills Training. *Prevention Science, 18*(4), 394–405.